

Using Galaxy for Metabolomics

ICG8

2013

Rob L Davidson PhD

Copyright NBAF-B 2013

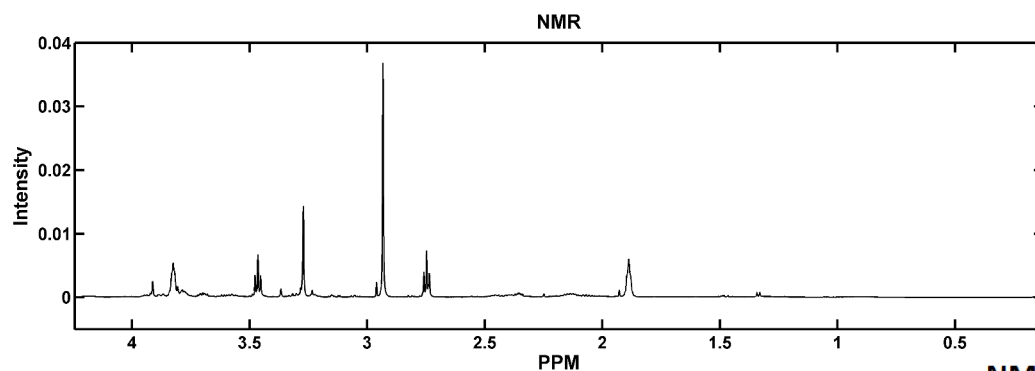


Overview



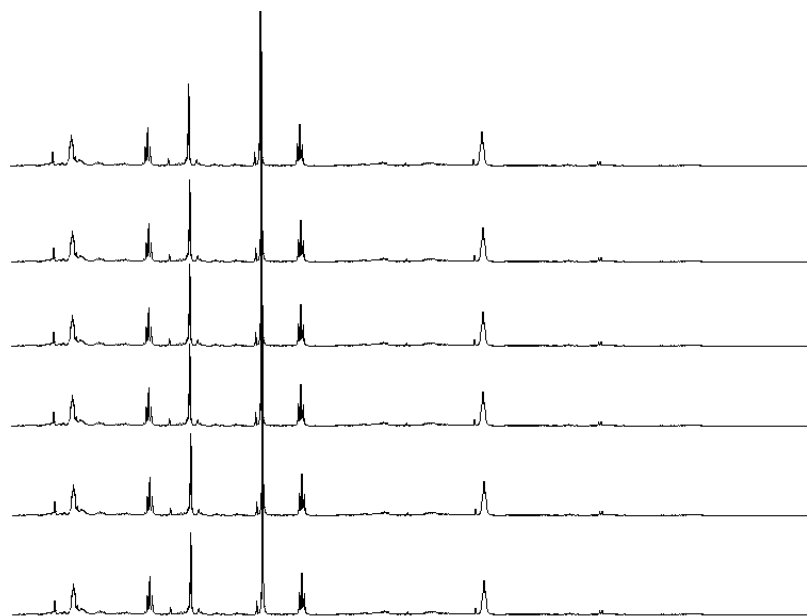
- **Metabolomics**
- **Galaxy**
- **Birmingham Workflow**
- **Galaxy Implementation**

Metabolomics - data



NMR spectrum = vector

	Peak 1	Peak 2	Peak3	Peak 4...	Peak N
Sample 1	19812	432	2309	4501882	5876



NMR spectra = matrix

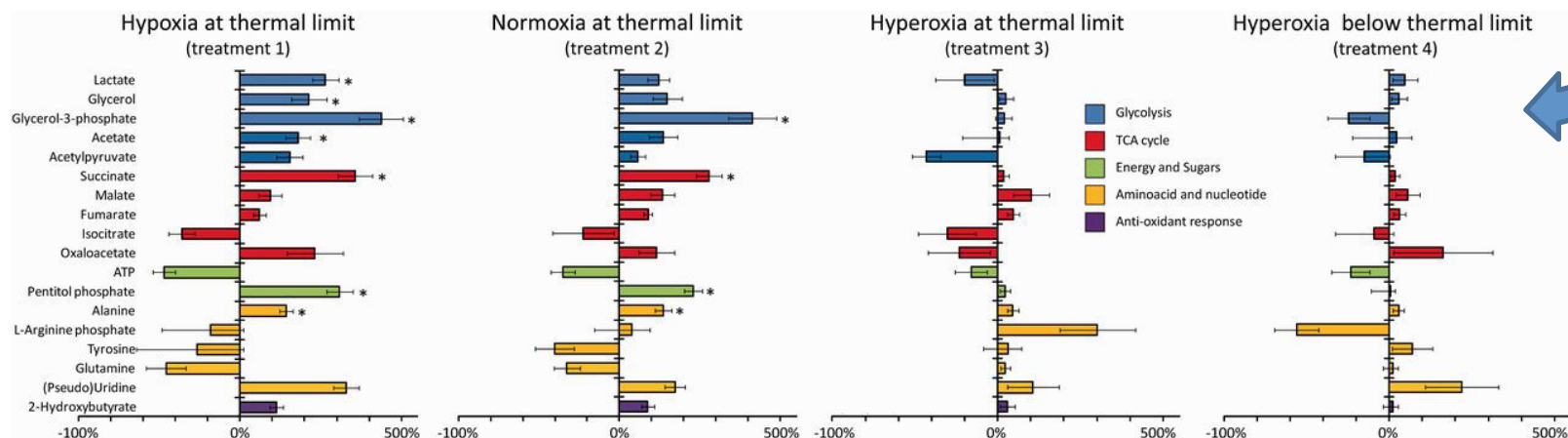
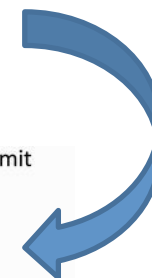
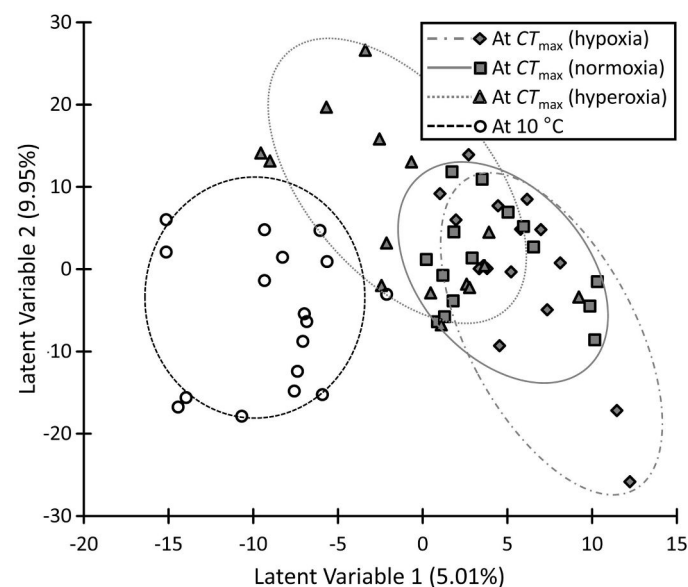
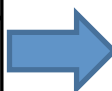
	Peak 1	Peak 2	Peak3	Peak 4...	Peak N
Sample 1	19812	432	2309	4501882	5876
Sample 2	8994	654	5409	357890	312
Sample 3	15012	1098	3102	1342098	10879
Sample 4	9999	302	4231	1809282	890
Sample N	17531	789	4500	2200192	3456

Metabolomics - tools

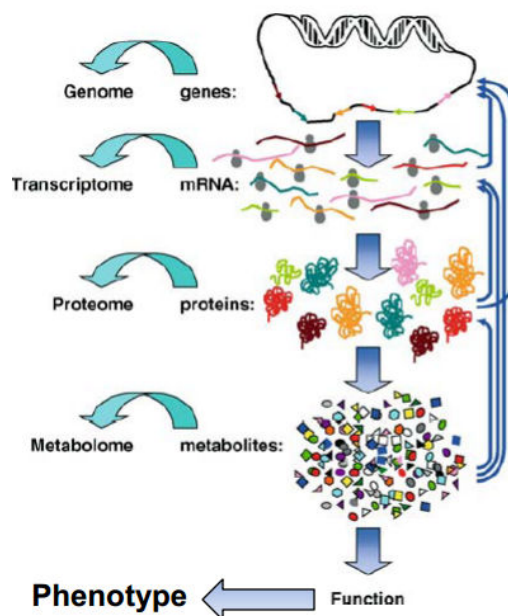
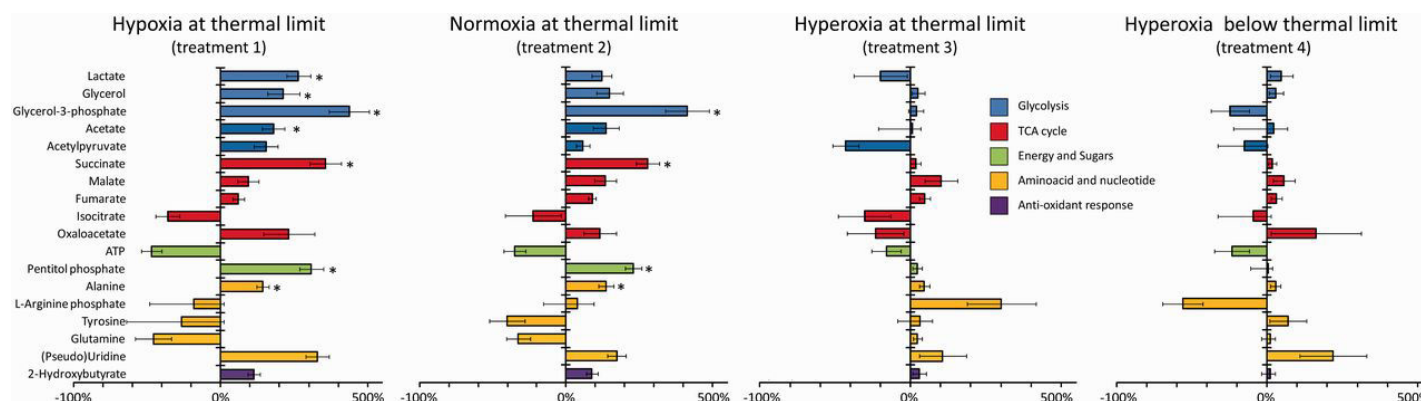


NMR spectra = matrix

	Peak 1	Peak 2	Peak3	Peak 4...	Peak N
Sample 1	19812	432	2309	4501882	5876
Sample 2	8994	654	5409	357890	312
Sample 3	15012	1098	3102	1342098	10879
Sample 4	9999	302	4231	1809282	890
Sample N	17531	789	4500	2200192	3456



Metabolomics - future



- Identify metabolite profile
- Link to pathways etc
- Correlate with other 'omes
- **Trans-omics!**

Galaxy

Copyright NBAF-B 2013



Galaxy



Galaxy

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy



Use the free public server

Get Galaxy



Install locally or in the cloud

Learn Galaxy



Screencasts, Galaxy 101, ...

Get Involved



Mailing lists, Tool Shed, wiki

Search all resources

The Galaxy Team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Open source

Over 20,000 main
Galaxy server users

Over 1200 papers
citing Galaxy use

45+ public
Galaxy servers

<http://galaxyproject.org>

Copyright NBAF-B 2013

Galaxy - interface



Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

The cluster on which many NGS tools run will be down for maintenance from 4 PM, Monday, Nov. 19 until 9 AM the following day (EST/EDT, UTC-0400). Jobs running on that cluster at that time

Tools

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

- Map with BWA for Illumina
- Map with BWA for SOLiD

ILLUMINA

- Map with Bowtie for Illumina

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Variant Detection

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

NGS: Picard (beta)

BEDTools

snpEff

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Map with BWA for Illumina (version 1.2.3)

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Select a reference genome:

Arabidopsis lyrata: Araly1

Is this library mate-paired?:

Single-end

FASTQ file:

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use:

Commonly Used

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

Suppress the header in the output SAM file:

☐

BWA produces SAM with several lines of header information

Execute

What it does
BWA is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (large), such as the human reference genome. It is developed by Heng Li at the Sanger Institute. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

Know what you are doing

There is no such thing (yet) as an automated gearshift in short read mapping. It is all like stick-shift driving in San Francisco. In other words = running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.

History

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Tool list

Tool parameterisation

Results panel

Galaxy - toolshed



Galaxy Tool Shed

Repositories Help User

Galaxy Tool Shed

Repositories

- [Browse by category](#)
- [Browse all repositories](#)
- [Login to create a repository](#)

Categories

Name	Description	Repositories
Assembly	Tools for working with assemblies	10
Computational chemistry	Tools for use in computational chemistry	2
Convert Formats	Tools for converting data formats	7
Data Source	Tools for retrieving data from external data sources	2
Fasta Manipulation	Tools for manipulating fasta data	13
Graphics	Tools producing images	4
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	12
Ontology Manipulation	Tools for manipulating ontologies	2
SAM	Tools for manipulating alignments in the SAM format	3
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	29
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	2
Statistics	Tools for generating statistics	4
Text Manipulation	Tools for manipulating data	9
Visualization	Tools for visualizing data	4

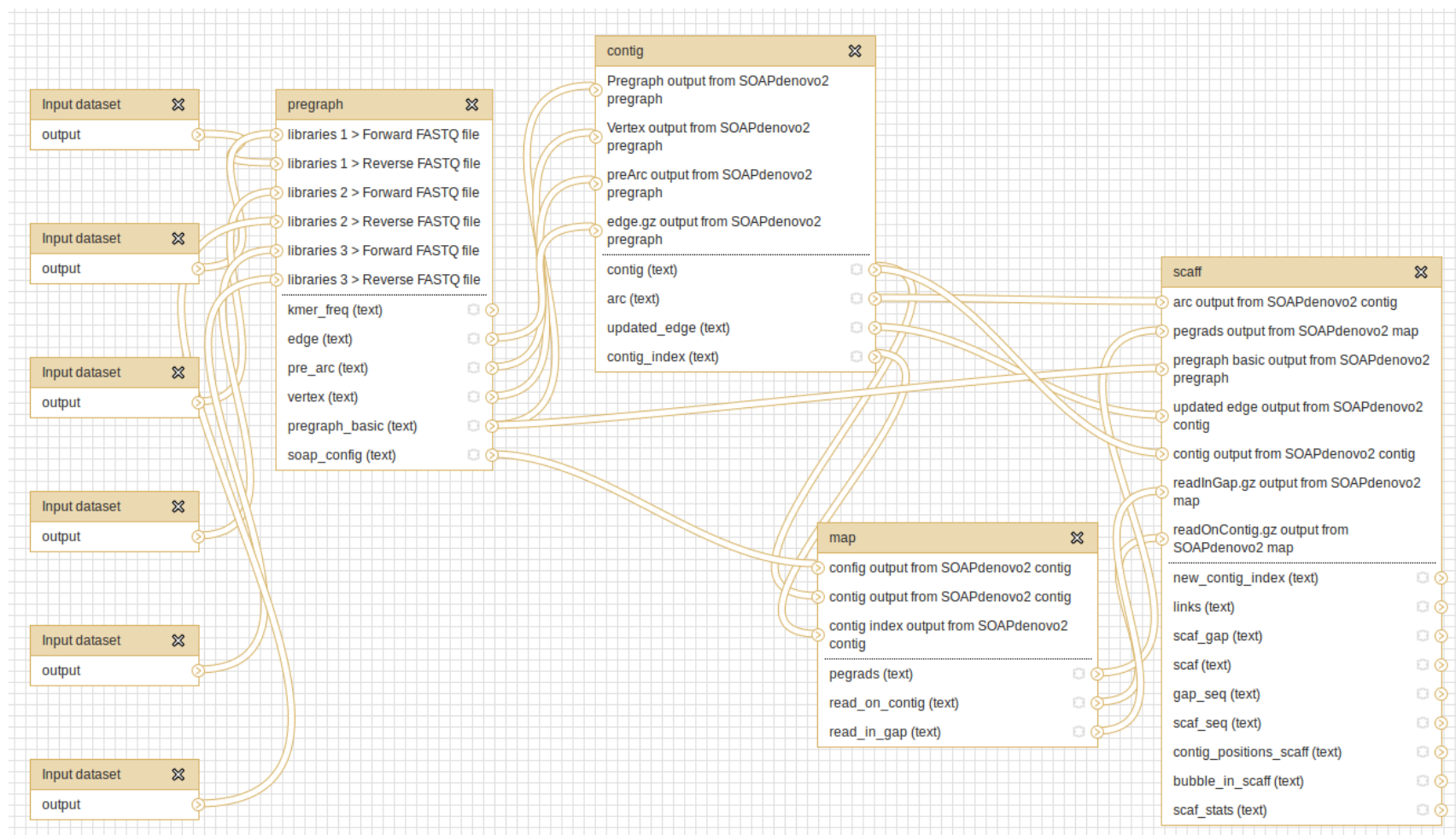
- Lots of tools ->
- Good for trans-omics
- But no metabolomics

...yet

<http://toolshed.g2.bx.psu.edu/>

Copyright NBAF-B 2013

Galaxy - workflows

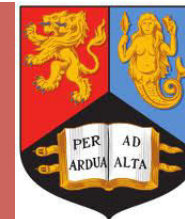


Metabolomics Workflow


Copyright NBAF-B 2013



Metabolomics Workflow

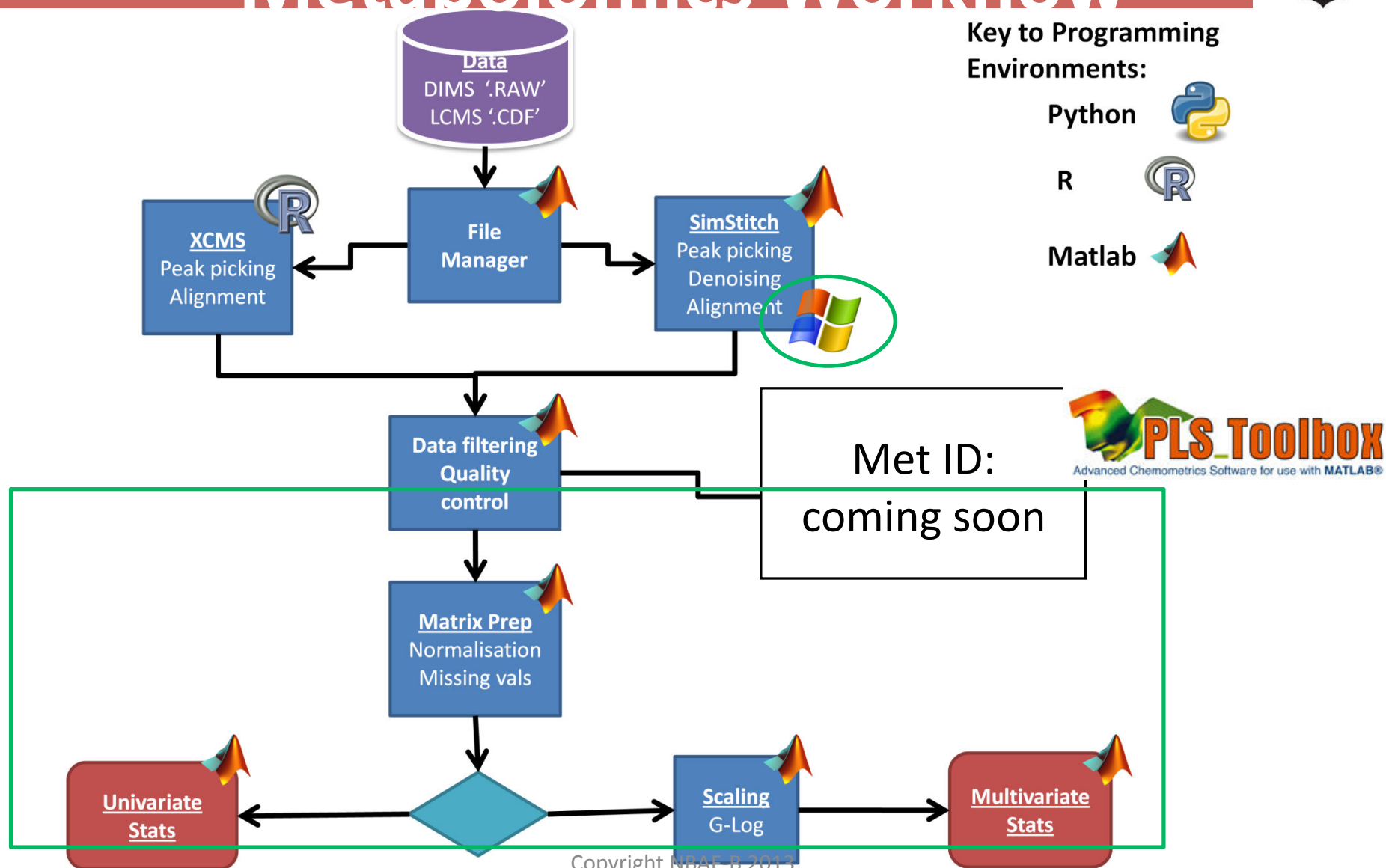


Key to Programming
Environments:

Python 

R 

Matlab 

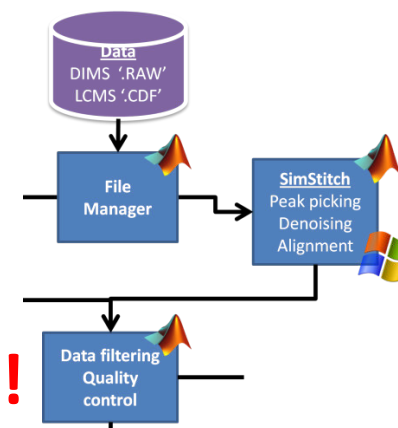


Metabolomics Workflow



- **SimStitch**

- FTICR-DIMS data
- RAW format ; **NB: requires MS Windows!**
- Stitches together short MZ ranges for greater accuracy
- Picks peaks/removes noise, aligns samples
- Applies filters to technical replicates, blanks and samples for greater robustness

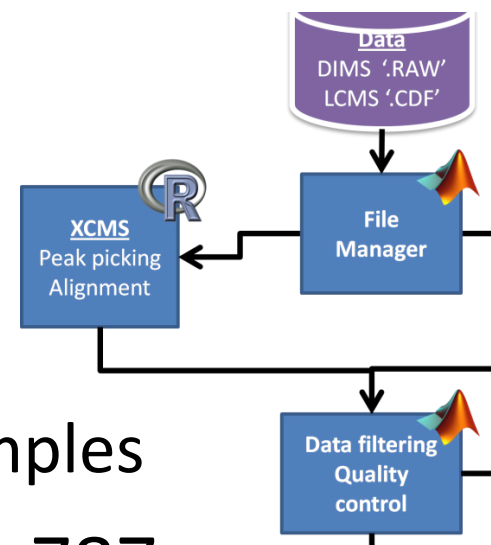


- *Southam AD, Anal Chem. 2007;79(12):4595-602.*

Metabolomics Workflow



- **XCMS**
 - LC-MS (also does GC-MS etc)
 - netCDF format (also does MZML etc)
 - Picks peaks/removes noise, aligns samples
- Smith CA, Anal. Chem. 2006 78:779-787
- <http://www.bioconductor.org/packages/2.12/bioc/html/xcms.html>
- In our pipeline, we call XCMS (in R) using Matlab...
 - To make use of our FileManager Structure
 - Because we also use Matlab for post-XCMS processing



Metabolomics Workflow



- **Matrix Prep:**

- PQN Normalisation

- Dieterle F, Anal Chem. 2006;78(13):4281-90.

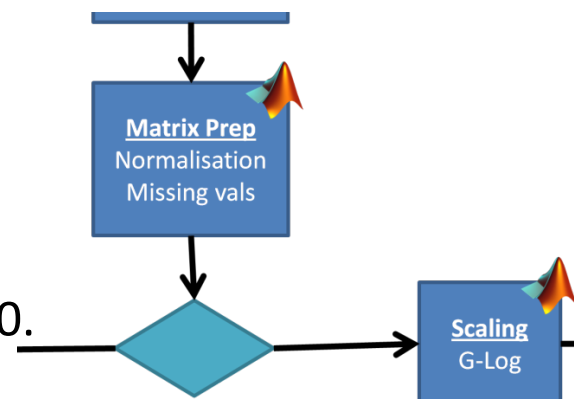
- KNN-Missing Value Imputation

- Hrydziuszko O, Metabolomics 2012 8(1):161-174

- G-Log scaling and variance stabilisation:

- Parsons HM, BMC Bioinf. 2007 8:234

- All done using PLS Toolbox data structures in Matlab

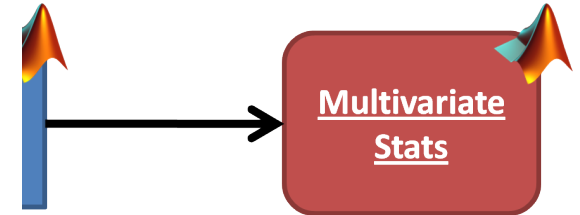


Metabolomics Workflow



- **Multivariate stats**

- PCA with automatic selection of PCs
- 2 classes: T-Test on scores on each PC
- 3+ classes: ANOVA and Tukey-Kramer
- **Output = text file containing these statistics**
- a PLS Toolbox 'model' is also created and scores plots etc can be viewed in Matlab



Metabolomics Workflow



- **Univariate stats**

- 2 classes: T-Test or Mann-Whitney-U for each peak
- 3+ classes: ANOVA or Kruskal-Wallis
- False Discovery Rate correction (Benjamini Hochberg)
- **Output = csv file containing these statistics**



Metabolomics Workflow



- **Done**
 - 1st end-end metabolomics pipeline in Galaxy
 - FTICR-DIMS and LCMS data
- **To Do**
 - Add in MI-Pack (underway)
 - Add more stats e.g. PLSDA
 - possibly merging with Netherlands Metabolomics <http://galaxy.nmcdsp.org/> (stats only)
 - Replace input file structure with ISA-Tab
 - <http://www.ebi.ac.uk/metabolights/>

Metabolomics Workflow

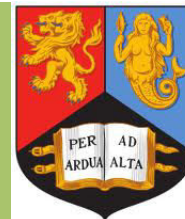


- **Requires 32bit MS Windows**
- **Data is large (100s Gb per study)**
- **Lots of processing power**
- **Multiple licenses**

Galaxy Implementation



Standard Galaxy

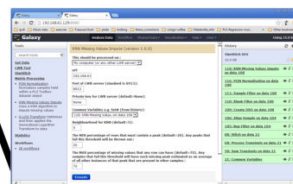


User's
Desktop PC



BUT!

Galaxy
Server



R Tool



Matlab Tool



Pythor



Metabo - Galaxy



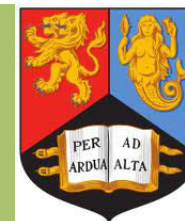
- Requirements
 - Allow Galaxy access to MS Windows
 - for FTICR-DIMS RAW file processing
 - Avoid passing large data over LAN
 - slow
 - Minimise cost of Galaxy implementation
 - make use of existing processors, storage and licences

Metabo - Galaxy



- Solution
 - Use Galaxy's Light Weight Runner (LWR)
 - Install LWR client on user's desktop
 - Adjust Python wrappers to send tools via LWR
 - Run all tools on User's desktop (MS Windows)
 - No need for
 - extra licenses
 - central storage/file transfer
 - powerful server

Metabo - Galaxy



Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

Tools

search tools

XCMS

SimStitch

- Common Variables Variables used by all/many SimStitch functions
- Sum Transients Step1 of SimStitch. Sums transient data
- Process Transients Step2 of SimStitch. Processes transient data
- Stitch Step3 of SimStitch. Stitches transient data
- Rep Filter Step4 of SimStitch. Filters and merges replicates.
- Align Sample Step 5 of SimStitch. Align sample spectra
- Create DSO Step 6 of SimStitch. Puts merged replicate spectra into PLS Toolbox DSO object.
- Blank Filter Step 7 of SimStitch. Filters out peaks that are heavily present in blank spectra.
- Missing Value Filter Step7 of SimStitch. Filter samples that have too many missing values.
- Sample Filter Step 8 of SimStitch. Filter peaks based on

Stitch (version 1.0.0)

This should be processed on:
My computer (or any other LWR server)

url:
10.0.2.2

Port of LWR server (standard is 8913):
8913

Private key for LWR server (default=None):
None

Common Variables e.g. Set# (from history):
[]

Noise Filter, to measure and filter noise (default checked):
☒

Include Noise Peaks in output and flag them? (default checked = yes):
☒

Min SNR (peak must have a height > MIN_SNR*(standard deviation of noise)):
10.0

Remove known regions of noise (default checked = yes):
☒

Known regions of high noise. Must be in pairs (start and end MZ) and separated by commas.:
101.6, 102.1, 1

Internally recalibrate (checked=default) or not (uncheck) where possible:
☒

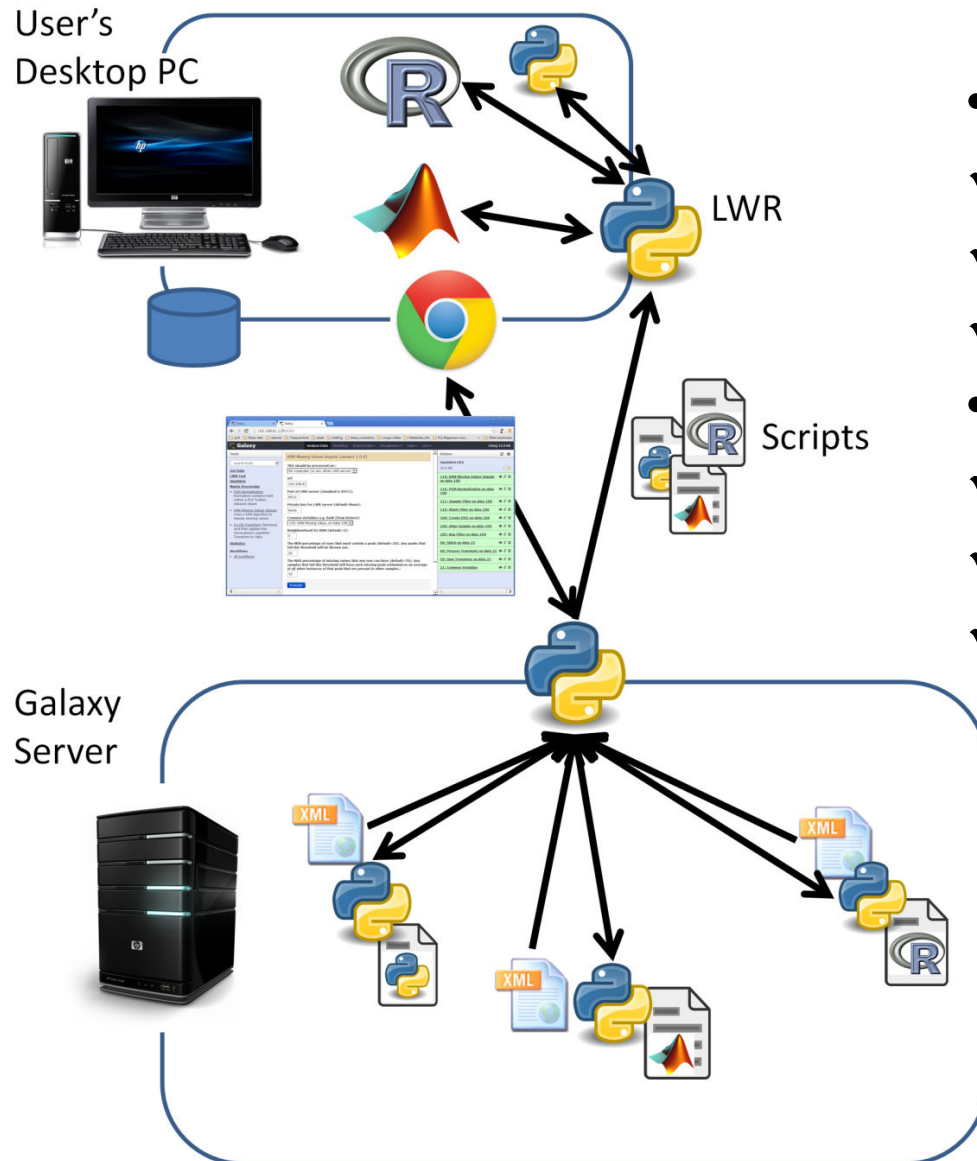
History

Unnamed history
0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

- New Input type
- Pre-fills user's IP
- Uses REMOTE_ADDR header

Metabo - Galaxy




- **Makes use of**
 - ✓ existing proc. power
 - ✓ licenses
 - ✓ user's MS Windows (RAW)
- **Still acts as**
 - ✓ version control
 - ✓ workflow manager
 - ✓ GUI

Metabo - Galaxy



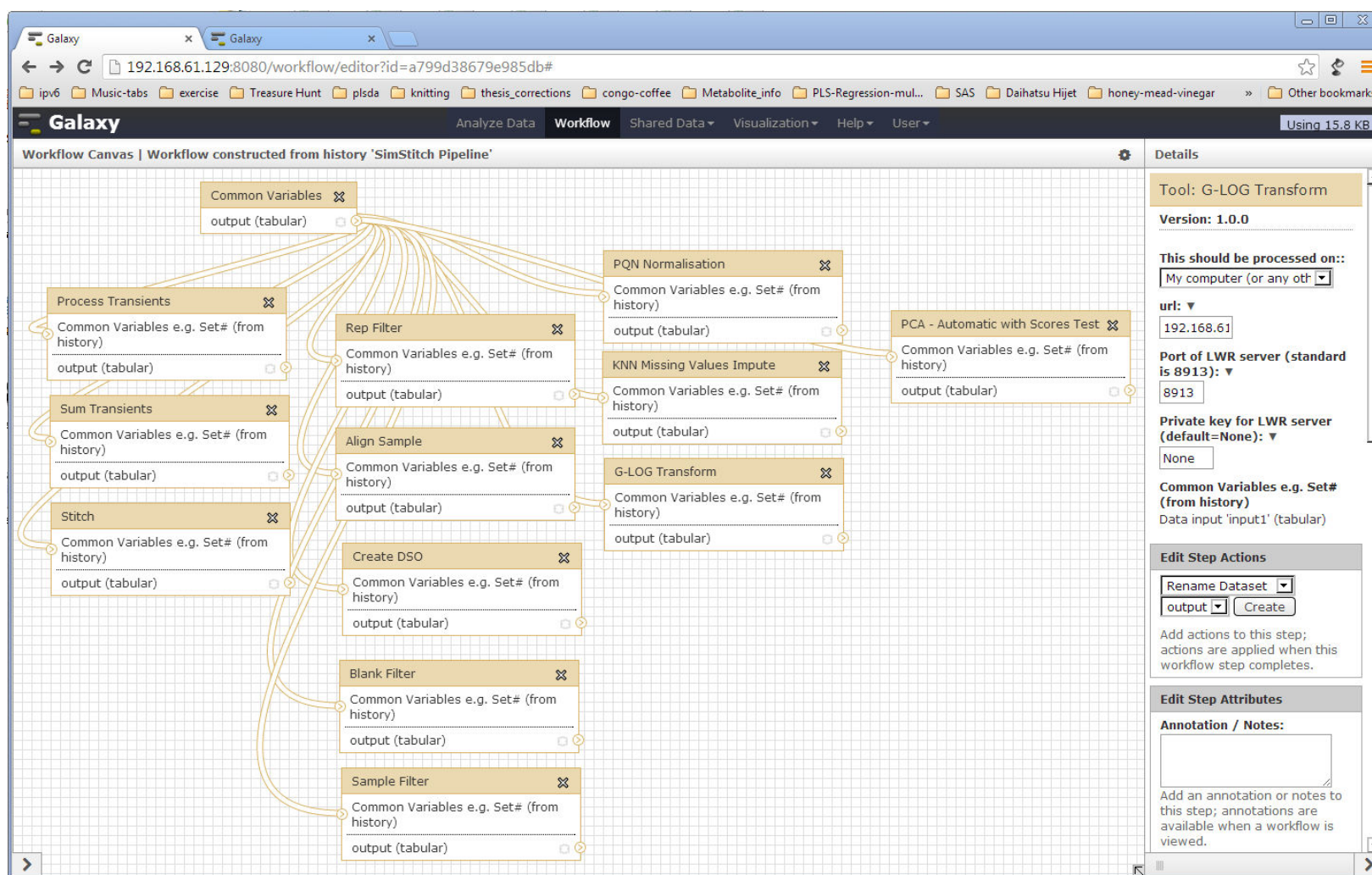
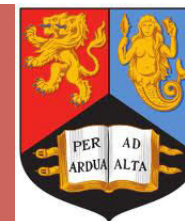
That said...

- Working with GigaScience 
 - <http://www.gigasciencejournal.com/>
- To be hosted on GigaGalaxy
 - <http://galaxy.cbiit.cuhk.edu.hk/>
- Using normal setup
 - (all processing/licenses on Galaxy server)
- Downloadable version to include both options

Summary



First RAW -> stats Galaxy Pipe



Summary



- Metabolomics has entered Galaxy!
- Can be expanded BY COMMUNITY!
- Can merge more easily with other 'omics
- Have developed Galaxy in a new way that allows
 - Use of existing hardware
 - Use of existing licenses
 - Less slow transfer of large data

Acknowledgements



- University of Birmingham
 - Ralf Weber, Ulf Sommer, Mark Viant
- Gigascience
 - Pete Li
- NERC Discipline Hop scheme

End



Questions?

r.l.davidson@bham.ac.uk